Contrasting the Automatic Identification of Two Discourse Markers in Multiparty Dialogues

Andrei Popescu-Belis
ISSCO / TIM / ETI
University of Geneva
Bd. du Pont-d'Arve 40
1211 Geneva 4, Switzerland
andrei.popescu-belis@issco.unige.ch

Sandrine Zufferey
Department of Linguistics
University of Geneva
Rue de Candolle 2
1211 Geneva 4, Switzerland
sandrine.zufferey@lettres.unige.ch

Abstract

The identification of occurrences of *like* and *well* that serve as discourse markers (DMs) is a classification problem which is studied here on a corpus of dialogue transcripts with more than 4,000 occurrences of each item. Decision trees using item-specific lexical, prosodic, positional and sociolinguistic features are trained using the C4.5 method. The results demonstrate improvement over past experiments, reaching the same range as inter-annotator agreement scores. DM identification appears to benefit from item-specific classifiers, which perform better than general purpose ones, thanks to the differentiated use of lexical features.

1 Introduction

The identification of discourse markers (DMs) is an essential step in dialogue understanding, especially when the lexical items used as DMs are ambiguous. *Like* and *well* are two frequent lexical items and potential DMs, which are among the most difficult ones to disambiguate, and they will serve here as a case study for automatic DM identification. The task will be discussed first from a linguistic and computational point of view. Previous attempts will be summarized, followed by the data, features and classifiers used here. The results will be discussed first by comparing our highest scores with baseline ones, then by analyzing the relevance to DM identification of various features. The best performances are shown to be comparable to inter-annotator agreement scores and higher than state-of-the-art scores, and lexical collocations are shown to be the most relevant features.

2 The Discourse Markers *Like* and *Well*

Despite the wide research interest raised by DMs for many years, there is no generally accepted definition or list of DMs (Andersen, 2001; Schiffrin, 1987). Items typically featured in this class are also called discourse connectives, pragmatic markers, or cue phrases, and include words and expressions such as *actually*, *and*, *but*, *I mean*, *like*, *so*, *you know* and *well*, which "generally have little lexical import but serve significant pragmatic functions in conversation" (Andersen, 2001, page 39).

For comparison purposes, we focus here on two lexical items, *like* and *well*, in order to determine the surface features that are most relevant to DM classifiers based on machine learning. These two items are among the most frequent and the most ambiguous DMs. *Like*, for instance, can be a preposition or an adverb, a verb or even a noun. When used as a DM, *like* is in reality much more than a filler, and can be more precisely described as a loose talk marker, signalling reported speech or an imprecise formulation of the speaker's belief, as in "He was like, yeah, I can make dogs raise their ears" or "It took, like, twenty minutes"—for more examples, see (Popescu-Belis and Zufferey, 2006, pages 7–9).

Well can also fulfil a variety of pragmatic and non-pragmatic functions (Schourup, 2001). When it is not a DM, *well* can be an adverb or an adjective ("He sings well", "I am well"), or a noun or verb ('water source'). As a DM, *well* can introduce a rejection of a previous request, or a disagreement with a previ-

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 10–17,
Antwerp, September 2007.

ous utterance, or can more generally mark hesitation or turn-taking, as in "Well, actually, you don't even need to do that..." or "Oh, yes, but well, uh, yes, but what I mean is that...".

3 Evaluation of DM Identification

The automatic identification of DMs is a binary classification task over the entire set of occurrences of the lexical item. Its evaluation requires a ground truth classification, and metrics to compare a candidate classification to it. The simplest evaluation metric is accuracy, i.e. the percentage of correctly classified instances (CCIs or C below). In addition, if DM identification is seen as the retrieval of the DMs among all occurrences of a lexical item, then recall (r), precision (p) and their f-measure (f) can be used to assess performance in a more detailed manner.

However, given that the distribution of DM vs. non-DM occurrences of a lexical item is seldom uniform, the above metrics should be corrected for chance agreement. To our knowledge, there are no widely used chance-corrected versions of recall and precision—the Kullback-Leibler divergence is seldom used for classification tasks—but a well-known agreement metric that *is* chance-corrected is the *kappa* (κ) score (Carletta, 1996). Although designed to measure inter-annotator agreement, κ quantifies the resemblance of two classifications by factoring out agreement by chance.

The κ score measures classification performance between -1 and 1, with random classification scoring 0. The κ measure is quite strict as it was designed to be sensitive to even small differences between coders. Therefore, a κ value above 0.67 is often considered a sign of acceptable agreement, while a value above 0.8 is considered very significant. According to Landis and Koch (1977), strength of agreement is fair for $0.2 < \kappa \leq 0.4$, moderate for $0.4 < \kappa \leq 0.6$, substantial for $0.6 < \kappa \leq 0.8$ and almost perfect above. In any case, the inter-coder agreement for the gold standard data represents the upper bound that can be legitimately expected from a classifier: even a perfect one cannot get closer to the gold standard than the humans who defined this standard.

4 Previous Studies of DM Identification

DMs play a considerable role in discourse processing tasks. For instance, some studies use discourse connectives to infer discourse structure (Reichman, 1985; Grosz and Sidner, 1986; Marcu, 1998), while others use DMs as cue words for discourse segmentation (Passonneau and Litman, 1997).

Many DMs, especially connectives or cue words, are not as highly ambiguous as *like* or *well*. Hutchinson (2004, page 686), for instance, targeted mainly the problem of automatic categorization of the pragmatic functions of discourse connectives, but only acknowledged the potential ambiguity of *and*. Similarly, Marcu's (1998) algorithm for DM identification, in relation to rhetorical parsing of written texts, aims at a list of 450 potential DMs, but *and* and *or* are ignored in many cases due to their ambiguity. It is also likely that *like* and *well* did not appear often in Marcu's 7200-word test data, over which recall was 80.8% and precision 89.5%.

Several studies have explicitly tackled DM identification in speech. Hirschberg and Litman (1993) applied a model based on intonational information to 34 DM types, and correctly classified 75.4% of their 878 classifiable tokens. Another model correctly classified 80.1% of the tokens based on human transcript and punctuation.

Siegel and McKeown (1994) proposed another transcript-based method, using decision tree classifiers constructed by a genetic algorithm, on a superset of the above data with 1,027 tokens. An interesting baseline score was obtained by a binary decision tree based only on the utterance-initial feature, which reaches 79.16% accuracy. The score of the best decision tree found by the genetic algorithm was only 79.20%. Although they did not improve performance over baseline, decision trees "discovered" some meaningful linguistic rules.

The relevance of machine learning techniques to DM identification was further emphasized by Litman (1996) in a set of experiments that extended and completed earlier studies by improving manually-derived classification models, using the same data set (34 DM types, 878 tokens). Litman used the C4.5 decision tree learner as well as an algorithm constructing sets of conditional rules. The features included prosodic features assigned by human an-

notators, textual features extracted from human transcripts, including correct punctuation, part of speech information assigned automatically, and the nature of the token itself. Most of the prosodic and textual models that were learned automatically outperformed corresponding models defined *a priori* by humans. The best performance using all available features was 16.9% error rate (83.1% accuracy) on the whole set.

DM identification was coupled to speech recognition, utterance segmentation, POS tagging, and repair correction by Heeman and Allen (1999). The best results are 97.26% recall, 96.32% precision, and 6.43% error rate, which was not, however, computed in the same sense as in the previous studies.

Comparison across studies is made difficult by the fact that the exact list of DMs differs from one study to another. In our study, only two DMs are contrasted, but they appear to be particularly multifunctional, hence difficult to disambiguate.

5 Description of the Data

The ratio between the number of targeted DM types (30–40) and the amount of data (often around 1,000 tokens) used in the previous studies did not allow for in-depth analysis of each DM, especially when a unique model was learned for all DMs. All studies except Heeman's were based on a monologue transcript (75 minutes, ca. 12,500 words), which was annotated by one or two linguists. Heeman's studies used transcripts from the TRAINS dialogue corpus, which contained 8,278 DMs among ca. 60,000 words. However, the exact list of DM types is not available (23 appeared as examples), nor the number of annotators or their agreement.

The data used here enables a more detailed study of *like* and *well* as a much larger number of occurrences is available. We use the ICSI Meeting Recorder Corpus of multi-party conversations, comprising transcripts of 75 meeting recordings with five to eight speakers (Janin et al., 2003). The meetings feature scientific discussions involving both native and non-native English speakers (52 in all). A distributional study and the *a posteriori* feature analysis show that there is no qualitative difference in the use of the two DMs by native vs. non-native but fluent speakers (Popescu-Belis and Zufferey, 2006, 6.3).

The recordings have a total duration of about 80 hours, corresponding to nearly 800,000 words in transcription. The segmentation into about 100,000 individual utterances is also available together with automatically generated word-level timing, based on forced alignment of transcript with audio, as well as indications of interruptions and unfinished utterances (Shriberg et al., 2004).

For this study, the DM and non-DM occurrences of the lexical items *like* and *well* were annotated by the two authors, with access to the dialogue transcripts and audio. In an experiment involving four non-expert annotators (Zufferey and Popescu-Belis, 2004), the observed inter-annotator agreement was $\kappa = 0.74$, but agreement between experts was not tested systematically. There are 4,519 occurrences of *like*, of which 2,052 (45%) serve as DMs, and 4,136 occurrences of *well*, of which 3,639 (88%) serve as DMs.

6 Features Used for DM Identification

The present method focuses on surface features only, since deeper analyses of an utterance seem to require in most cases the prior identification of DMs. For instance, it would not be realistic to assume the availability of a parse tree or of a deep semantic analysis of an utterance, as their construction would precisely require knowledge of DMs. However, joint models for POS tagging or parsing with DM identification could incorporate knowledge about DMs as presented here.

Lexical features model the words immediately preceding or following a DM candidate, and depend on the width of the lexical window ($2N$) and the minimal frequency (F) of words used as possible values. One feature is defined for each position with respect to the DM candidate: WORD($-N$), ..., WORD(-1), WORD($+1$), ..., WORD($+N$). The possible values of these variables are the words observed around the DM candidates, above a certain frequency F, or 'other', or 'none' if there is no such position in the utterance (this implicitly includes information about the candidate's position). For a window of width $N = 1$, i.e. using only WORD(-1) and WORD($+1$), the frequency thresholds of $F = 3$, $F = 10$ and $F = 20$ correspond respectively to 360, 150 and 90 word types as possible values.

DMs also have specific **positional and prosodic** properties, but not all the prosodic features are easy to extract automatically. The following features, derived from the forced-alignment segmentation at the word level and the ground truth segmentation into utterances, will be used: INITIAL: set to 'yes' if the candidate is the first word of an utterance, to 'no' otherwise; FINAL: set to 'yes-completed' if the candidate is the last word of a completed utterance, to 'yes-interrupted' if it is the last word of an interrupted utterance and to 'no' otherwise; PAUSE-BEFORE: the duration of the pause before the candidate, or 10 seconds if the utterance starts with it; PAUSE-AFTER: the duration of the pause after the candidate, or 10 seconds if it ends the utterance; DURATION: the duration of the candidate. The first two are positional features, while the latter three are very elementary prosodic or temporal features.

The following **speaker-related, sociolinguistic** features will also be used, with the following possible values: GENDER: 'female' or 'male'; AGE: an integer; EDUCATION: 'undergraduate', 'graduate', 'PhD', 'professor'; NATIVE: 'native' vs. 'non-native' English speaker; ORIGIN: 'UK', 'US East', 'US West', 'US other', and 'other'. Such features could be useful to a dialogue processing system that is used frequently by the same persons.

For each category, the features were selected based on potential linguistic and computational relevance. In addition, the TYPE feature represents the nature of the candidate DM, either *like* or *well*, allowing the two lexical items to be processed differently, as in (Litman, 1996).

7 DM Classifiers

The choice of a classifier for DM identification is constrained by the nature of the features: some are discrete while others are continuous; the lexical features are quite sparse and have an unclear impact on classification. Here, four types of classifiers were tested using the WEKA toolkit (Witten and Frank, 2000): Bayesian Networks (BN), Support Vector Machines (SVM), decision trees, and k-nearest neighbours ($k = 3$), which performed below the first three, so its results are not reported here.

Decision tree classifiers are made of nodes that test features of a DM-candidate, and of branches that correspond to the possible values of the features. Each terminal node is labelled with one of the two classes, DM or non-DM (Siegel and McKeown, 1994; Litman, 1996). Decision trees can be learned from training data using the C4.5 method (Quinlan, 1993), which accepts both discrete and continuous features. C4.5 constructs a nearly optimal decision tree classifier for the training data, that is, a tree that maximizes the number of correctly classified instances (CCIs) over the training data, but not necessarily recall, precision or *kappa*.

8 DM Identification Results

The best scores reached by the classifiers do not differ substantially in our experiments, as the 95%-confidence intervals computed using 10-fold cross-validation (training on 90% of the data and testing on 10%) are not disjoint. The best scores are obtained by a Bayesian Network that uses only the discrete features of the DMs—see first line of Table 1. Decision trees will be used preferentially below, as BN classifiers take longer to build and are more difficult to interpret than them, and their performance is only slightly higher.

8.1 Highest Scores vs. Baseline Scores

Baseline scores for DM identification are at least 50% CCIs because of the binary nature of the classification problem. As shown in the last three lines of Table 2, the majority classifier, which assigns to all candidates the type of the most frequent class observed in the training data reaches scores that are well above zero for at least three metrics out of five. Only κ appears to be insensitive to this bias.

Method	Test	CCIs (%)	κ	r	p	f
MAJ	*l*+*w*	65.75	0	.66	1	.79
	l	45.40	0	1	.45	.62
	w	87.99	0	1	.88	.94
ISM	*l*+*w*	70.55	.42	.64	.88	.74
	l	54.60	0	0	0	0
	w	87.98	0	1	.88	.94

Table 2: Baseline scores for the majority classifier (MAJ) and for an item-specific majority classifier (ISM), tested on *like* and *well* together (noted *l*+*w*), then separately for each item.

Method	Train	Test	CCIs (%)	κ	r	p	f
BN	*l*+*w*	*l*+*w*	**90.480**±.646	**.783**±.016	.957±.004	.904±.008	**.930**±.005
	l+*w*	*l*	84.009±1.431	.681±.028	.896±.012	.784±.021	.836±.014
	l+*w*	*w*	97.537±.456	.880±.021	.991±.004	.981±.005	.986±.003
SVM	*l*+*w*	*l*+*w*	**89.290**±.571	**.752**±.014	.964±.006	.884±.008	**.922**±.004
	l+*w*	*l*	82.908±1.216	.661±.023	.914±.016	.759±.020	.829±.013
	l+*w*	*w*	96.250±.841	.808±.037	.992±.005	.966±.009	.979±.005
C4.5	*l*+*w*	*l*+*w*	**88.862**±.511	**.751**±.011	.923±.007	.909±.006	**.916**±.004
	l+*w*	*l*	81.046±.885	.618±.018	.802±.020	.785±.013	.793±.013
	l+*w*	*w*	97.396±.443	.870±.026	.991±.002	.980±.005	.985±.002

Table 1: Best results obtained by three machine learning algorithms, trained and tested on *like* and *well* together (noted *l*+*w*), and then also tested separately on each item (noted *l* and respectively *w*). The three most significant metrics (scores in **bold**) yield clearly decreasing scores from the first to the third condition.

The use of the TYPE feature, allowing an item-specific majority classifier to distinguish between the lexical items *like* and *well*, increases the baseline scores (see ISM in Table 2). This classifier, based only on the following rules: "*like* is not a DM" and "*well* is a DM", reaches already $\kappa = 0.42$.

The scores of the four classifiers from Table 1 are significantly above the baseline. The fact that the best score is $\kappa = 0.78$ shows that automatic DM identification performances are in the same range as human inter-annotator agreement. The best scores are also higher than those obtained by the classifiers that use only a subset of features, as shown in the next sub-section.

The scores of the best BN classifier applied separately to *like* and *well* are also shown in Table 1, 2nd and 3rd lines. These are significantly higher for the identification of DM *well* ($\kappa = 0.880, f = 0.986$) than for DM *like* ($\kappa = 0.681, f = 0.836$). It is true that *well* as a DM is much more frequent than *like* as a DM (ca. 88% vs. 45%), so the baseline accuracy is higher for *well* (CCI = 88% vs. CCI = 45%, see 2nd and 3rd lines of Table 2) but this effect should be filtered out at least by the κ metric—nevertheless, which is still much higher for *well* than for *like*. *Well* appears thus to be easier to identify than *like*, with the features used here.

8.2 Relevance of the Features

The best-scoring decision tree uses four **lexical features** (WORD(-2), WORD(-1), WORD($+1$) and WORD($+2$)), their possible values being all the word types occurring at least 10 times in this 4-word lexical window ($F = 10, N = 2$). The best C4.5 learner was set to construct binary trees with at least two instances per leaf.

Four experiments were particularly informative. First, using only the WORD(-1) lexical feature, i.e. the lexical item preceding the candidate DM, C4.5 constructs trees that contain at the uppermost node the lexical collocations that are the most reliable indicators of a DM, with scores reaching CCI $= 86.5\%, \kappa = 0.68, r = 0.97, p = 0.85, f = 0.90$, which are not much below the best possible ones. When distinguishing *like* from *well* in the decision trees, thanks to the TYPE feature in addition to WORD(-1), the scores increase to CCI $= 87.4\%, \kappa = 0.72, r = 0.91, p = 0.90, f = 0.90$ (note the high value of κ).

Words situated after the candidate DM appear to be much less informative: if only TYPE and WORD($+1$) are used, CCI $= 77.8\%$ and $\kappa = 0.47$. When all lexical features encoded as WORD(n) are used ($n \leq 2$), the results are getting even closer to the best ones, but recall increases and precision decreases. The lexical features, and in particular the word before the candidate, appear thus to be nearly sufficient for DM identification of *like* and *well*. The actual values of WORD(n) that serve as lexical indicators are not, of course, the same for the two items.

Turning now to **positional and prosodic features**, experiments using combinations of one, two or three features are summarized in Table 3. A first series of experiments with positional features (left

Positional						Prosodic / temporal					
Features	CCIs(%)	κ	r	p	f	Features	CCIs(%)	κ	r	p	f
T	70.5	0.42	0.64	0.88	0.74	T	70.5	0.42	0.64	0.88	0.74
I	68.8	0.42	0.54	0.97	0.70	B	74.2	0.50	0.65	0.94	0.77
T+I	73.4	0.46	0.70	0.87	0.78	T+B	75.3	0.48	0.75	0.86	0.80
F	67.5	0.09	0.98	0.67	0.80	A	67.5	0.09	0.98	0.67	0.80
T+F	72.5	0.46	0.64	0.91	0.75	T+A	75.8	0.50	0.74	0.87	0.80
T+I+F	75.8	0.51	0.71	0.90	0.79	T+A+B	79.4	0.55	0.82	0.86	0.84

Table 3: Results with C4.5 decision trees using combinations of positional and prosodic / temporal features (T: TYPE, I: INITIAL, F: FINAL, B: PAUSE-BEFORE, A: PAUSE-AFTER).

part of the table) shows that on average, classification is improved as more features become available among the following: TYPE (T), INITIAL (I), and FINAL (F). These results are paralleled by a second series (right column), obtained with prosodic/temporal features, PAUSE-BEFORE (B) and PAUSE-AFTER (A), in which scores also increase when more features are available. As the second series uses features that implicitly encode more information than in the first one, superior results are obtained. The best decision trees using PAUSE-BEFORE contain the following rule: "*like* is a DM only when the pause before it is longer than 0.06 s", indicating that a pause approximately longer than 60 milliseconds is a good indicator of a DM. A similar value (though with a lower score) is found for the pause after DM *like*, while no effect was observed for *well*. In addition, other experiments have shown that DURATION is not a relevant feature. Prosodic features appear thus to be superior to positional ones, but remain inferior to lexical features.

The **sociolinguistic features** alone do not permit the construction of a classifier with a non-zero score if the two lexical items *like* and *well* are not distinguished. When they are, the best decision tree generated by C4.5 remains the majority classifier for *well* ("all occurrences are DMs") and a more refined classifier for *like*: a number of heavy DM-*like* users are identified, for which all occurrences of *like* that they produce are considered as DMs. The scores of this classifier are: $\mathrm{CCI} = 77.3\%, \kappa = 0.47, r = 0.88, p = 0.80, f = 0.84$. These values are clearly above the scores obtained using TYPE only.

A number of sociolinguistic features appear to be relevant in the case of *like* only (the baseline score being here $\kappa = 0$). Using EDUCATION, the best tree found by C4.5 reaches $\kappa = 0.39$ with the following rule: "if the speaker is an undergraduate or a graduate, consider all tokens of *like* as DMs; if the speaker is a post-doc or a professor, consider all tokens of *like* as non-DMs". A similar correlation ($\kappa = 0.40$) holds for the region of ORIGIN ('US West' implies heavy DM *like* user) and a stronger correlation ($\kappa = 0.44$) holds for AGE ('under 30' implies heavy DM user). These experiments thus bring statistical evidence that younger speakers from the US West tend to overuse *like* as a DM, which corroborates a view commonly held by sociolinguists, who often consider the DM *like* as a feature of adolescent speech (Andersen, 2001). Since in our data there were a majority of speakers under 30 from the US West, below PhD level, it is not possible to identify the precise feature that correlates with DM-*like* overuse among AGE, ORIGIN or EDUCATION)—more subjects are needed to "decorrelate" these features, though the present number (52) is sufficient to explore each feature in part.

8.3 Automatic Attribute Selection

Two methods were used to compare the merits of features, and appear to lead to similar results. WEKA's correlation-based feature subset selection algorithm (CFS) aims at finding the best subset of features by examining the individual predictive power of each feature and at the same time minimizing redundancy within the subset. Alternatively, independent relevance scores for each feature can be computed using two criteria: the information gain and χ^2 (Witten and Frank, 2000). Their rankings being very similar, only information gain is used here.

Like		*Well*	
Feature	IG	Feature	IG
WORD(−1)	.44	WORD(−1)	.39
WORD(+1)	.21	PAUSE-BEFORE	.23
SPEAKER	.15	INITIAL	.19
PAUSE-BEFORE	.06	WORD(+1)	.15
AGE	.06	PAUSE-AFTER	.10
PAUSE-AFTER	.05	FINAL	.10
EDUCATION	.04	SPEAKER	.04
INITIAL	.03	DURATION	.03
COUNTRY	.02	AGE	.01
FINAL	.01	COUNTRY	.005
GENDER	.01	EDUCATION	.004
DURATION	.01	NATIVE	.001
NATIVE	.001	GENDER	.001

Table 4: Separate ranking of features for *like* and *well* according to their information gain (IG). Significant IG decreases are indicated by a line.

The CFS algorithm finds the following optimal subset of attributes: {TYPE, PAUSE-BEFORE, INITIAL, WORD(−1)}, thus confirming previous observations. The word before the candidate is a key feature, along with the specific processing of each DM (TYPE), and the pause before the candidate (or its utterance-initial character).

The ranking of each feature shows that the most distinctive feature is the word before the candidate, WORD(−1), followed at some distance by PAUSE-BEFORE, INITIAL, WORD(+1) (the word after the candidate) and TYPE. The ranking can also be done separately with respect to *like* and *well*, as shown in Table 4. The lists are similar to the one just described for the joint identification task.

Attribute selection can also be used to determine the most discriminative collocations, i.e. the words that best indicate whether their neighbouring candidate is likely to be a DM or not DM (words must be used individually as features in this case). The best feature set found by CFS for *like* contains *something, things, seems* (if they precede *like*, then the occurrence is not a DM), or *that* (if it follows *like*, then the occurrence is not a DM). Similar trials focused only on *well* help to determine collocations such as *very well, as well, how well*, which are relevant to identify non-DM occurrences of *well*.

9 Discussion

To summarize, the best scores for *like* and *well* are: $\mathrm{CCI} = 90\%, \kappa = 0.78, r = 0.96, p = 0.90, f = 0.93$, obtained for a Bayesian Network; the best scores of a C4.5 decision tree or an SVM are not much lower. These scores are well above the baseline ones, although this depends on how the baseline is defined, as some very simple classifiers have scores that are well above zero. These scores also compare favourably to the ones obtained in previous studies, although the DMs and evaluation measures sometimes differ considerably.

The best scores obtained are comparable to inter-annotator agreement values observed for non-expert subjects ($\kappa = 0.74$). This indicates that automatic classifiers may have reached the highest possible performance in the present experiments, and that the set of features was sufficient to reach an accuracy comparable to human annotators. Improving the scores seems thus to require also a more reliable annotation, obtained for instance by allowing experienced annotators to discuss and to adjudicate their individual annotations.

The most important features appear to be the lexical collocations that can be learned from the training data. Among these, the word before a candidate DM is the most useful one, especially as it implicitly encodes also the utterance-initial character. Scores obtained using only lexical features are within 5% distance from the best overall scores. Decision trees based only on lexical features, or even on TYPE and WORD(−1) only, are not far from optimal ones. It is therefore surprising that these features were not used in Litman's (1996) study, maybe from lack of enough training data for each item.

Positional and prosodic features are significantly less efficient than lexical ones, when used alone, although they appear in the best decision trees just below lexical features. The sociolinguistic features are only slightly correlated to DM use, almost exclusively for *like*: the most reliable indicators are the identity and the age/education of the speakers.

The TYPE feature is crucial: *like* and *well* are much better processed separately than as a unique class. This conclusion confirms, on a large data set, the theoretical insights arguing that DMs are not a homogeneous class. Although some of the pre-

vious features do generalize to both lexical items (such as PAUSE-BEFORE), many of the features are item-specific, as found also by Litman (1996), and in particular the lexical features, which appeared to be the most relevant ones. Overall, this study has shown that DM identification can reach accuracies that are comparable to inter-annotator agreement scores, if item-specific classifiers using lexical features are trained on large corpora.

Future work should focus first on the application of the method to other ambiguous DM candidates, such as *you know*. This requires, for each item, the manual annotation of a sizeable amount of instances for training and test, and possibly some adaptation of the features. More elaborate prosodic features should also be studied. Finally, DM classifiers could be applied prior to POS tagging and parsing, or could be integrated into POS taggers or parsers.

Acknowledgments

This work has been supported by the Swiss National Science Foundation through the IM2 NCCR on Interactive Multimodal Information Management.

References

Gisle Andersen. 2001. *Pragmatic Markers of Sociolinguistic Variation*. John Benjamins, Amsterdam.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Barbara J. Grosz and Candace L. Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Peter A. Heeman and James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers utterances in spoken dialogue. *Computational Linguistics*, 25(4):1–45.

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Ben Hutchinson. 2004. Acquiring the meaning of discourse markers. In *Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*, pages 685–692, Barcelona, Spain.

Adam Janin, Don Baron, Jane A. Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of ICASSP 2003 (IEEE International Conference on Acoustics, Speech, and Signal Processing)*, Hong Kong, China.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Diane J. Litman. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

Daniel Marcu. 1998. A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. In *ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 1–7, Montreal, Canada.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–140.

Andrei Popescu-Belis and Sandrine Zufferey. 2006. Automatic identification of discourse markers in multiparty dialogues. Working paper 65, ISSCO, University of Geneva, December 2006.

John R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Francisco, CA.

Rachel Reichman. 1985. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model)*. MIT Press, Cambridge, MA.

Deborah Schiffrin. 1987. *Discourse Markers*. Cambridge University Press, Cambridge, UK.

Lawrence C. Schourup. 2001. Rethinking 'well'. *Journal of Pragmatics*, 33:1025–1060.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue)*, pages 97–100, Cambridge, MA.

Eric V. Siegel and Kathleen R. McKeown. 1994. Emergent linguistic rules from inducing decision trees: Disambiguating discourse clue words. In *Proceedings of AAAI 1994 (12th National Conference on Artificial Intelligence)*, pages 820–826, Seattle, WA.

Iain Witten and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, CA.

Sandrine Zufferey and Andrei Popescu-Belis. 2004. Towards automatic identification of discourse markers in dialogs: The case of like. In *Proceedings of SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue)*, pages 63–71, Cambridge, MA.